ON THE USE OF NONLOCAL AND NON POSITIVE DEFINITE BASIS FUNCTIONS IN RADIAL BASIS FUNCTION NETWORKS

D Lowe

Aston University, UK

ABSTRACT

It is invariably the case that when an application is developed using the Radial Basis Function network in the neural network domain, it is constructed using Gaussian basis functions. This paper discusses the rationale for employing alternative basis functions to the prevalent Gaussian. In particular we argue the case in support of unbounded basis functions and non positive definite basis functions. The use of unbounded and nonpositive basis functions, though counterintuitive in application domains such as classification and time series forecasting, have a good theoretical motivation from the domains of functional interpolation and somewhat surprisingly from kernel based density estimation. In addition to collating the theoretical arguments, we present a performance comparison between Gaussian and unbounded, non positive definite basis functions in a Radial Basis Function network applied to a financial derivatives regression problem: estimating the price of $/DM options contracts.

INTRODUCTION

The Radial Basis Function Network is a very simple, flexible and interpretable structure which may be employed in classification/discrimination tasks, time series prediction and other mapping tasks and also in an unsupervised manner to perform a type of topographic feature extraction. The basic network structure may be derived and motivated from a variety of perspectives. In particular its classification and discrimination abilities may be understood from a statistical pattern processing perspective [5, 12]. In contrast its ability to perform time series prediction and other mapping problems are related to a deterministic dynamical systems viewpoint and functional interpolation. Its original transposition into the neural network domain [2] was from this latter interpretation. Both perspectives had the unifying philosophical basis that the aim of the network is to approximate the underlying structure which generated the observed data, rather than the data itself.

In spite of its simplicity and interpretational power, the practical exploitation of the Radial Basis Function network invariably employs the ubiquitous Gaussian as the nonlinear basis function. This paper puts the theoretical case for employing alternative basis functions in practical problems, and demonstrates its use on a real world problem taken from the financial domain.

KERNEL FUNCTIONS, PROBABILITIES AND APPROXIMATION

We have already commented that the Radial Basis Function network may be derived from a statistical pattern processing background. In particular in classification problems under certain assumptions the optimum output of the network $y_c(\boldsymbol{x}) = \sum_{j=0}^{m} \lambda_{jc}\phi_j(||\boldsymbol{x}-\mu_j||)$ attempts to approximate the class conditional distribution of the data $p(c|\boldsymbol{x})$ (c is a specific class and $\boldsymbol{x}$ is a vector of patterns input to the network) whereas the hidden layer of basis functions may be optimised to approximate the unconditional distribution of the data, $p(\boldsymbol{x})$ [6] - the class information being provided by optimising the hidden-to-output layer weights. Thus the Radial Basis Function network may be interpreted from the perspective of kernel based density estimation methods. Of course if we require the network as a whole to approximate probabilities in the strict sense, one way of achieving this is to assume a Mixture model form and employ positive definite and normalisable basis functions which are themselves density functions. This might be sufficient but it is not necessary.

The theory of kernel based density estimation [3] produces many recommended bounded ker-

'Artificial Neural Networks', 26–28 June 1995, Conference Publication No. 409 © IEE 1995

nels which may or may not be density functions themselves. Examples of the basis functions $\phi(z)$ are the Epanechnikov ($\frac{3}{4}[1-z^2]$ for $|z| < 1$, and 0 otherwise) which is an optimal kernel in the sense of minimising asymptotic mean square error, the Biweight ($\frac{15}{16}[1-z^2]^2$ for $|z| < 1$ and 0 otherwise), and of course the Gaussian.

Biased density estimates

Now if we restrict the basis functions such that $\int \phi(x)\ dx = 1$ and $\phi(x) \geq 0\ \forall x$ then it is clear that there exists a Radial Basis Function network (with fixed final layer weights) which is also consistent with these restrictions. This could be considered an advantage if we wish to use the Radial Basis Function network as a non-parametric density estimator, i.e by constructing the network out of linear combinations of functions which themselves have the properties of probability density functions. However these constraints carry a hidden penalty in that the resulting density estimator is necessarily biased, and in addition the asymptotic convergence of the bias is limited. This follows from results of Rosenblatt [7], Yamato [13], Shapiro [11] and others [3].

In particular, if $g(x)$ is a kernel estimator of a probability density function $f(x)$ based upon a finite set of n independent and identically distributed random variables, such that the kernel functions $K(x)$ of the estimator satisifies (a) $K(x) \geq 0\ \forall x \in \mathbb{R}$ and (b) $\int_{\mathbf{R}} K(x)dx = 1$, then the kernel estimator itself is necessarily biased ($\langle g(x) \rangle \neq f(x)\ \ \forall x$). The result of Rosenblatt is slightly less specific in that if the estimator $g(x)$ based upon a finite set of samples is everywhere positive as a function of x, then it is a biased estimator. Consequently we see that any kernel-based density estimator constructed out of positive kernels and a finite set of samples cannot be unbiased.

However we know that in the asymptotic case the class of kernel estimators is capable of zero bias (implicity this is connected to the universality of Radial Basis Function networks). We can therefore enquire how the class of kernel estimators asymptotically reduces the bias. There are several results on the asymptotic convergence of kernel based density estimators. One example is due to Shapiro which indicates that if we allow the kernel functions to assume negative regions, then the asymptotic convergence of the bias is faster than if we insisted upon using positive definite basis functions.

Of course we are interested in reducing both the bias and the variance simultaneously if our estimator is a good one – an estimator with zero bias but with a high variance is not likely to be particularly effective. The asymptotic variance depends on $\lim_{n\to\infty} var[g(x)] = f(x)\int K^2(x)dx/(n\sigma)$. Therefore provided that we choose the 'smoothing parameter' σ of the kernel function in such a way that $lim_{n\to\infty}\sigma = 0$ and $lim_{n\to\infty} n\sigma = \infty$ then the variance should tend to zero if the kernel is square integrable (Parzen [9]).

In summary, there is no kernel estimator with non negative kernels which is unbiased for all continuous functions. So for the finite sample case we must tolerate a certain level of bias in our estimator. It is an interesting point that relaxing the positive definiteness restriction on the basis function allows a density estimator with potentially smaller finite sample bias (and zero bias in some instances), and in addition should have better convergence behaviour. This then is an argument provided from kernel based density estimation theory why it might be advantageous to use basis functions which are *not* themselves density functions and indeed may have negative regions, provided that the network as a whole has the desired behaviour.

Nonlocal basis functions and localised response

In addition to the basis functions suggested by kernel density estimation, functional approximation theory provides additional choices. From the theory of functional interpolation, the following choices of kernel, or basis function are common. Linear (z), Cubic (z^3), Thin Plate Spline ($z^2 \log z$), Inverse Multiquadric ($[z^2 + c^2]^{-1/2}$), Multiquadric ($[z^2 + c^2]^{1/2}$), and again the Gaussian. Note that these functions do not have finite support, and indeed some of the choices are *unbounded* functions, contrary to intuition and common folklore that the network basis functions are localised. However it is nevertheless correct that the parameters of the network may be chosen such that $y(x) \to 0$ as

$x \rightarrow \infty$ so that the network as a whole achieves a localised response, where $y(x)$ is the output of a (scalar) network. We present an intuitive illustration of this property later in the paper exploiting the concepts of the *dual basis space*, or the space of *equivalent smoothing kernels*. Actually a stronger result holds [10] in that under the conditions that the centres form a regular grid, then if the desired actual function to be approximated, $f(x)$ is a low order polynomial, then one of the conditions that must be satisfied if $y(x) \equiv f(x)$ is that $\int \phi(x)dx$ is unbounded (so Gaussians are not suitable for instance).

A final interesting property of unbounded basis functions, which is really related to the computation universality of Radial Basis Function networks [8] is the following result due to Brown (see the Appendix in [10] for details). If $f : \mathcal{D} \rightarrow \mathbb{R}$ is a continuous function which maps a closed, bounded submanifold, $\mathcal{D}$ of $\mathbb{R}^d$ into $\mathbb{R}$ that it is desired to approximate, then it is possible to prove that irrespective of the form of f there exists a finite set of centres μ_j such that $f(x)$ may be arbitrarily closely approximated by a Radial Basis Function with *unbounded* basis functions. Specifically $\left| f(x) - \sum_{j=1}^{n} \lambda_j \|x - \mu_j\| \right| < \epsilon$. Note that this is a rare result for a finite Radial Basis Function architecture in producing continuous mappings from a closed domain.

Although these are rather formal and extreme results, they do at least provide existence examples of why using *unbounded* basis functions can be advantageous. To confirm these formal results we first consider the case of unbounded basis functions combining to provide a *localised* response. We do this in terms of the *dual basis* space.

DUAL BASIS SPACE

There are two ways to represent a Radial Basis Function. The first is the 'usual' way in terms of a linear combination of 'fixed' (i.e. data–independent basis functions), such that the output of the k–th final layer node may be represented as

$$o_k(\boldsymbol{x}) = \sum_{j=0}^{m} \phi_j(\boldsymbol{x}) \lambda_{j,k}$$

for m basis functions, $\phi_j(\boldsymbol{x})$ which are conveniently thought of as being 'located' at 'centres' in the data space. In this representation,the basis function calculation is straightforward and most of the work goes into calculating the coefficients of the basis functions, i.e. the 'weights' $\lambda_{j,k}$. The second method of representation is in terms of the 'dual basis' functions which we can motivate as follows [1].

If the network weights are optimised according to a least squares approach, then analytically the optimum values of the weight matrix $\boldsymbol{\Lambda}$ is given in terms of the pseudo–inverse of the data matrix $\boldsymbol{\Phi}^+$ and the P training patterns in the target matrix as :– $\boldsymbol{\Lambda} = \boldsymbol{\Phi}^+ \boldsymbol{T}$. Thus, we can express the output of the network as

$$\begin{aligned} o_k(x) &= \sum_{j=0}^{m} \phi_j(\boldsymbol{x}) \sum_{l=1}^{P} [\boldsymbol{\Phi}^+]_{j,l} [\boldsymbol{T}]_{l,k} \\ &= \sum_{l=1}^{P} \left[\sum_{j=0}^{m} \phi_j(\boldsymbol{x}) [\boldsymbol{\Phi}^+]_{j,l} \right] [\boldsymbol{T}]_{l,k} \\ &= \sum_{l=1}^{P} \Psi_l(x) [\boldsymbol{T}]_{l,k} \end{aligned}$$

In this expansion the Radial Basis Function output is a linear combination of *new* basis functions, but now the weighting coefficients are given directly (by the target matrix). However the basis functions themselves are data–dependent $\Psi_l(x) = \sum_{j=0}^{m} \phi_j(\boldsymbol{x}) [\boldsymbol{\Phi}^+]_{j,l}$ and have to be evaluated. The computational load has shifted from evaluating a network with adjustable weights and fixed basis functions, to evaluating datum-specific kernel functions which are combined linearly by predetermined weights. Of course the network output is the same and there is no computational advantage, but it does allow an alternative interpretation of behaviour.

We now give an example where even if the original basis functions $\phi(\ldots)$ are *nonlocal*, by construction the dual basis functions $\Psi(\ldots)$ are *localised* around their appropriate centres. Since the output of the RBF is a finite linear combination of these 'dual basis' localised functions, we see that the overall response of the network may be interpreted as a localised reponse despite using nonlocalised functions in the first place.

[1] This is a trivial generalisation of the concept of *kernel smoothers* in standard regression by linear additive models as discussed in [4]

Simple Example: Noisy Sine Wave

Figure 1 shows data generated according to a noisy sine wave on $[0, 1]$, with the data density increasing quadratically towards smaller values of the argument. A simple regression experiment was constructed to 'discover' the underlying sine wave generator, but using a Radial Basis Function network with $z^2 \log(z)$ (i.e. non-local) basis functions. Six such basis functions were positioned uniformly on $[-1, 2]$. The predicted network mapping function superimposed on Figure 1 indicates that the regression experiment was successful. However we are not interested in the experiment itself, rather we are concerned with its interpretation How can we interpret the success of this experiment given that we are forming linear combinations of non-local basis functions? To answer this we need to look at the behaviour of the dual basis functions.

Figure 1: *Training data generated by a noisy sine wave superimposed on the final mapping function produced by the Radial Basis Function network using $z^2 \log(z)$ basis functions.*

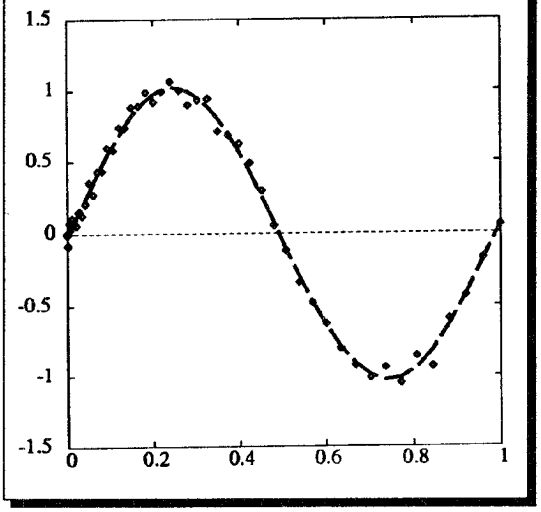
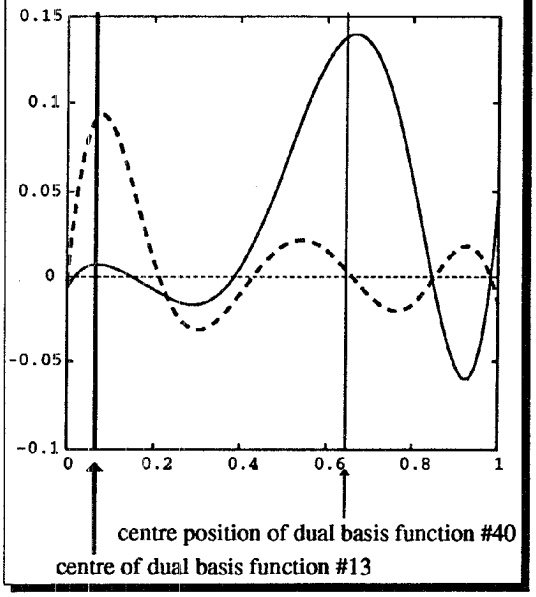


Figure 2 shows two of the basis functions (recall there is one dual basis function for every data point). Note two characteristics: one is that these basis functions exhibit localised (though damped oscillatory) behaviour [2] on the domain of the data , and secondly, dual basis functions in regions of higher data density tend to be more tightly localised. Hence a finite linear combination of such dual basis functions should also produce a well-behaved localised response, as the network mapping function in Figure 2 shows. This explains the behaviour of the experiment in a practical manner.

[2] However, note that outside the domain of the data the reponse is 'unregulated' since it corresponds to 'extrapolation'

Figure 2: *Variation of two of the constructed 'dual' basis functions. Solid line corresponds to the basis function centred at data point 40, dashed line centred at data point 13.*



EXPERIMENT 2: PRICING OF FINANCIAL OPTIONS CONTRACTS

Finally we discuss the behaviour of Radial Basis Function networks using Gaussian and thin plate spline functions, when applied to a more difficult real world problem. The example considered is a nonlinear multivariate regression problem based on the pricing of financial instruments known as *Options*. An option on a commodity is a contract which gives the owner of the contract the right (but not the obligation) to either buy or sell the underlying commodity at a predetermined price at some time in the future. The price in the contract is known as the *exercise price* or the *strike price*. The date in the contract is known as the *expiration date*, *exercise date* or *maturity*. A contract giving the owner the right to buy by a certain date is known as a 'Call', and the right to sell is known as a 'Put' options contract. If the contract can only be exchanged at the expiry date of the contract, then it is known as a *European* option. If

the contract can be exercised at any time up to the expiry date it is known as an *American* option. Options on stocks were first traded on an organised exchange in 1973, since when there has been a dramatic growth in the options markets. The underlying assets of the contracts now include stocks, stock indices, currencies, debt, commodities and futures contracts. Although often used for balancing risk by constructing portfolios, options can be used for speculation due to their inbuilt gearing of capital. When used to speculate it is possible to gain significantly and lose all the initial investment. Hence mispricing or incorrectly estimating the value of an options contract can have significant consequences, as the 1995 saga of the Baring Bank has emphatically illustrated.

This specific area of financial mathematics is particularly interesting, as there exists a widely accepted model which estimates the price of European call or put options by an analytic formula which involves five variables. The model is known as the Black–Scholes pricing formula and was derived in 1973 [1]. The model makes various assumptions, such as equilibrium markets, stock prices follow a random walk in continuous time and are lognormally distributed, there exist constant short term interest rates, and there are no dividends or transaction costs and also assuming a rational investor. Under these assumptions, Black and Scholes produced an analytic formula for the value of a European call or put option in terms of five quantities: the underlying stock price, S, the strike price X, the risk free interest rate r the time to expiry T and the 'volatility' σ (an estimate of the standard deviation of the logarithm of price fluctuations). The value of a European call option is predicted to be

$$C = SN(d_1) - Xe^{-r(T-t)}N(d_2)$$

where $d_1 = [ln(S/X) + (R + \sigma^2/2)(T - t)]/\sigma\sqrt{T-t}$ and $d_2 = d_1 - \sigma\sqrt{T-t}$, and $N(x)$ is the cumulative probability distribution function for a standardised normal variable. However this formula, though widely in use today, is known to consistently underestimate or overestimate the market depending upon whether the option contract is a call or a put. Despite many efforts to improve on the basic model, there is still no reliable model to estimate the market value of options contracts.

The following experiment used a Radial Basis Function with both Gaussian and thin plate spline basis functions to produce a regression model to estimate the market value of European call options on \$/DM exchange rates. Training data was taken from 42 consecutive days between February 8th 1989 and April 18th 1989. This amounted to 1592 separate contracts. The risk free interest rates were taken from the current values of 3 month Treasury Bills and volatility was calculated using an optimally chosen sliding window on historic data. All other information was easily available. The network performance was tested against the subsequent 21 consecutive trading days, between April 19th 1989 and May 23rd 1989, amounting to 514 different call contracts. The input data was prewhitened to be zero mean, unit variance on a channel by channel basis (with parameters evaluated in the training data only). This should be more consistent with the assumptions of using spherical Gaussian basis functons and hence we should expect reasonable performance using a Gaussian RBF network.

Figures 4 and 5 show the test set performance obtained by the two networks, using the same number of basis functions centred at the same locations, but one using localised and one using nonlocal basis functions. The scatterplots show the actual market value of the contracts plotted against the Radial Basis Function predictions. Clearly both models have produced very good estimates of the underlying values of the contracts. The Gaussian RBF fit has a Pearson rank order correlation of 0.9507 and the thin plate spline network has a rank order correlation of 0.9753, hence indicating very good correlations between the network models and the data. However one can note from the figures that the nonlocal basis function network tend to produce more robust estimators, particularly for outliers and high priced contracts where the data density is clearly more sparse. This experimental conclusion is consistent with the theoretical arguments discussed at the start of this paper.

Figure 3: *Test set results: Scatterplot of the RBF predicted vs the actual market prices of the $/DM Call option contracts using 15 Gaussian basis functions.*

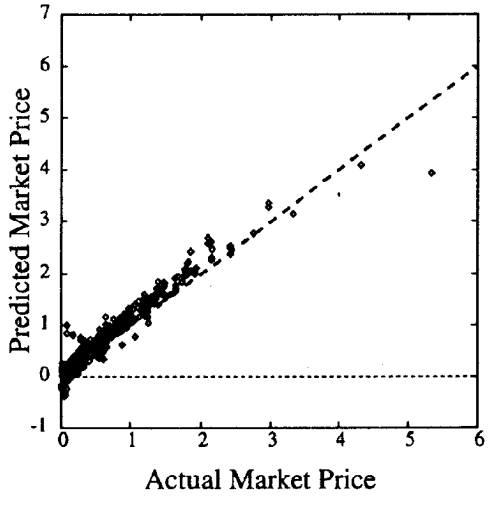
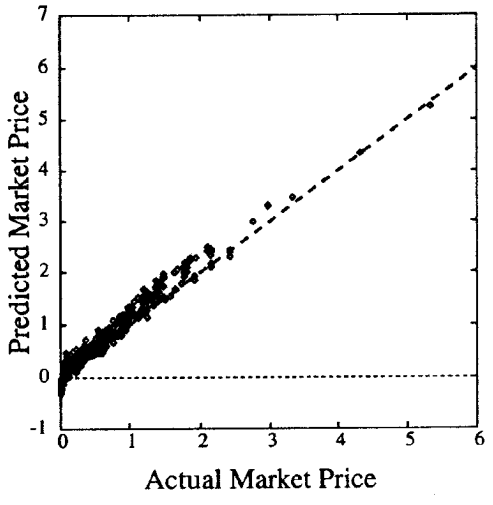


Figure 4: *Test set results: Scatterplot of the RBF predicted vs the actual market prices of the $/DM Call option contracts using 15 thin plate spline basis functions, positioned randomly at the same locations as the Gaussian basis functions of the previous Figure.*



Acknowledgements

The author would like to thank the EPSRC for support of this work under contract #GR/J75425.

References

[1] Black F, and Scholes M, "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, **81**, 637–659, (1973).

[2] Broomhead D S and Lowe D, (1988) "Multivariable Functional Interpolation and Adaptive Networks", *Complex Systems*, **2**(3),269–303.

[3] Hand D, (1982), "Kernel Discriminant Analysis", (Research Studies Press, John Wiley & Sons).

[4] Hastie, T J and Tibshirani, R J, (1990), "Generalised Additive Models", (Monographs on Statistics and Applied Probability 43, Chapman and Hall).

[5] Lowe D, (1991), "On the Iterative Inversion of RBF Networks: A Statistical Interpretation", 2nd *IEE International Conference on Artificial Neural Networks, Conference Publication number 349*, 29–33.

[6] Lowe D, (1995), "Radial Basis Functions", in *The Handbook of Brain Theory and Neural Networks*, Michael A. Arbib (ed), Bradford Books/MIT Press.

[7] Rosenblatt M, "Remarks on some nonparametric estimates of a density function" *Ann. Math. Statist.*,**27**,832–837, (1956).

[8] Park J and Sandberg I W, "Universal approximation using Radial Basis Function networks", *Neural Computation*, **3**, 246–257, (1991).

[9] Parzen E, "On estimation of a probability density function and mode", *Ann Math. Statistics*, **33**, 1065–1076, (1962).

[10] Powell M J D, (1990), "The theory of Radial Basis Function approximation in 1990", DAMTP preprint 1990/NA11.

[11] Shapiro, J S, "Smoothing and Approximation of Functions", Van Nostrand Reinhold, New York, (1969).

[12] Tråvén H G C, "A Neural Network Approach to Statistical Pattern Classification by "Semiparametric" Estimation of Probability Density Functions", *IEEE Transactions on Neural Networks*, **2**(3), 366–377, (1991).

[13] Yamato H, "Some statistical properties of estimators of density and distribution functions", *Bull. Math. Statistics*, **19**,113–131, (1972).